SYSTEMS AND METHODS FOR CLASSIFYING AUDIO INTO BROAD PHONEME CLASSES

GOVERNMENT CONTRACT

[0001] The U.S. Government may have a paid-up license in this invention and the right in limited circumstances to require the patent owner to license others on reasonable terms as provided for by the terms of Contract No. DARPA F30602-97-C-0253.

RELATED APPLICATIONS

[0002] This application claims priority under 35 U.S.C. § 119 based on U.S. Provisional Application No. 60/419,214 filed October 17, 2002, the disclosure of which is incorporated herein by reference.

BACKGROUND OF THE INVENTION

A. Field of the Invention

[0003] The present invention relates generally to speech processing and, more particularly, to audio classification.

B. Description of Related Art

[0004] Speech has not traditionally been valued as an archival information source. As effective as the spoken word is for communicating, archiving spoken segments in a useful and easily retrievable manner has long been a difficult proposition. Although the act of recording audio is not difficult, automatically transcribing and indexing speech in an intelligent and useful manner can be difficult.

[0005] Speech is typically received by a speech recognition system as a continuous stream of words without breaks. In order to effectively use the speech in information management systems (e.g., information retrieval, natural language processing and real-time alerting systems), the speech recognition system initially processes the speech to generate a formatted version of the speech. The speech may be transcribed and linguistic information, such as sentence structures, may be associated with the transcription.

[0006] Additionally, information relating to the speakers may be associated with the transcription.

[0007] Speech recognition systems, when transcribing audio that contains speech, may classify the audio into a number of different audio classifications. The audio classifications may include classifications, such as speech/non-speech and vowel/consonant portions of the audio. The speech recognition system may use the speech classifications when processing the speech signals. Non-speech regions, for example, are not transcribed. Also, whether a vowel or consonant is being spoken may dictate which acoustic model to use in analyzing the audio.

[0008] Conventional speech recognition systems may decode an incoming audio stream into a series of phonemes, where a phoneme is the smallest acoustic event that distinguishes one word from another. The phonemes may then be used to classify the audio signal. The number of phonemes used to represent a particular language may vary depending on the particular phoneme model that is employed. For English, a complete phoneme set may include approximately 50 different phones.

[0009] One problem associated with generating a complete phoneme set for an incoming audio signal is that the complete phoneme set may be computationally burdensome, particularly when attempting to process speech in real-time.

[0010] Thus, there is a need in the art to more efficiently classify segments of an audio signal.

SUMMARY OF THE INVENTION

[0011] Systems and methods consistent with the principles of this invention classify audio into broad classes. The broad nature of the classes allows for quick classification. The audio classifications may then be used in performing more detailed speech recognition functions.

[0012] One aspect of the invention is directed to a method for classifying an audio signal containing speech information. The method includes receiving the audio signal and classifying a sound in the audio signal as a vowel class when a first phoneme-based model determines that the sound corresponds to a sound represented by a set of phonemes that define vowels. The method also includes classifying the sound in the audio signal as a fricative class when a second phoneme-based model determines that the sound corresponds to a sound represented by a set of phonemes that define consonants. Further, the method includes classifying the sound in the audio signal based on at least one non-phoneme based model.

[0013] A second aspect of the invention is directed to a method of training audio classification models. The method comprises receiving a training audio signal and receiving phoneme classes corresponding to the training audio signal. A first Hidden Markov Model (HMM) is trained based on the training audio signal and the phoneme classes. The first HMM classifies speech as belonging to a vowel class when the first HMM determines that the speech corresponds to a sound represented by a set of phonemes that define vowels. A second HMM is trained based on the training audio signal and the phoneme classes. The second HMM classifies speech as belonging to a fricative class when the second HMM determines that the speech corresponds to a sound represented by a set of phonemes that define consonants.

[0014] Another aspect of the invention is directed to an audio classification device. The audio classification device includes a signal analysis component that receives an audio signal and processes the audio signal by converting the audio signal to the frequency domain and/or generating cepstral features for the audio signal. The device further includes a decoder that classifies portions of the audio signal as belonging to classes. The classes include a first phoneme-based class that applies to the audio signal when a portion of the audio signal corresponds to a sound represented by a set of phonemes that define vowels. A second phoneme-based class applies to the audio signal when a portion of the audio signal corresponds to a sound represented by a set of phonemes that define consonants. The classes further include at least one non-phoneme class.

[0015] A system consistent with another aspect of the invention includes an indexer, a memory system, and a server. The indexer receives input audio data and generates a rich transcription from the audio data. The indexer includes audio classification logic that classifies the input audio data into at least one of a number of broad audio classes, where the broad audio classes include a phoneme-based vowel class, a phoneme-based fricative class, a non-phoneme based bandwidth class, and a non-phoneme based gender class. The indexer further includes a speech recognition component that generates the rich transcriptions based on the broad audio classes determined by the audio classification logic. The memory system stores the rich transcriptions. The server receives requests for documents and responds to the requests by transmitting one or more of the rich transcriptions that match the requests.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate the invention and, together with the description, explain the invention. In the drawings,

[0017] Fig. 1 is a diagram of a system in which systems and methods consistent with the present invention may be implemented;

[0018] Fig. 2 is a diagram illustrating an exemplary computing device;

[0019] Fig. 3 is a diagram illustrating functional components of the audio classification logic shown in Fig. 1;

[0020] Fig. 4 is a diagram illustrating the decoder shown in Fig. 3 in additional detail;

[0021] Fig. 5 is a diagram illustrating an exemplary sequence of output classification symbols;

[0022] Fig. 6 is a diagram illustrating training of a phoneme based model consistent with an implementation of the invention; and

[0023] Fig. 7 is a diagram illustrating training of a phoneme based model consistent with another implementation of the invention.

DETAILED DESCRIPTION

[0024] The following detailed description of the invention refers to the accompanying drawings. The same reference numbers in different drawings may identify the same or similar elements. Also, the following detailed description does not limit the invention. Instead, the scope of the invention is defined by the appended claims and equivalents.

[0025] An audio classification system classifies sounds in an audio stream as belonging to one of a relatively small number of classes. In one implementation, seven classes are used: vowels, fricatives, narrowband, wideband, coughing, gender, and silence. The classified audio may be used to enhance speech recognition of the audio.

EXEMPLARY SYSTEM

[0026] Fig. 1 is a diagram of an exemplary system 100 in which systems and methods consistent with the present invention may be implemented. In general, system 100 provides indexing and retrieval of input audio for clients 150. For example, system 100 may index input speech data, create a structural summarization of the data, and provide tools for searching and browsing the stored data.

[0027] System 100 may include multimedia sources 110, an indexer 120, memory system 130, and server 140 connected to clients 150 via network 160. Network 160 may include any type of network, such as a local area network (LAN), a wide area network (WAN) (e.g., the Internet), a public telephone network (e.g., the Public Switched Telephone Network (PSTN)), a virtual private network (VPN), or a combination of networks. The various connections shown in Fig. 1 may be made via wired, wireless, and/or optical connections.

[0028] Multimedia sources 110 may include one or more audio sources, such as radio broadcasts, or video sources, such as television broadcasts. Indexer 120 may receive audio data from one of these sources as an audio stream or file.

[0029] Indexer 120 may receive the input audio data from multimedia sources 110 and generate a rich transcription therefrom. For example, indexer 120 may segment the input data by speaker, cluster audio segments from the same speaker, identify speakers by name or gender, and transcribe the spoken words. Indexer 120 may also segment the input data based on topic and locate the names of people, places, and organizations. Indexer 120 may further analyze the input data to identify when each word was spoken (possibly based on a time value). Indexer 120 may include any or all of this information as metadata associated with the transcription of the input audio data. To this end, indexer 120 may include audio classification logic 121, speaker segmentation logic 122, speech recognition logic 123, speaker clustering logic 124, speaker identification logic 125, name spotting logic 126, topic classification logic 127, and story segmentation logic 128.

[0030] Audio classification logic 121 initially classifies the audio from multimedia sources 110 into one of a number of broad classes (e.g., 7) in a manner consistent with the present invention. As previously mentioned, the audio classes may include: vowels, fricatives, narrowband, wideband, coughing, gender, and silence. The classified audio may then be used by logic 122-128 as each of these elements perform their individual functions. Audio classification logic 121 will be discussed in more detail below.

[0031] Speaker segmentation logic 122 detects changes in speakers. Speaker segmentation logic 122 may use the classes determined by audio classification logic 121. For example, changes in the gender of a speaker may indicate a speaker change.

[0032] Speech recognition logic 123 may use statistical models, such as acoustic models and language models, to process input audio data. Different acoustic and language models may be trained based on the different classes generated by audio classification logic 121. The language models may include n-gram language models, where the probability of each word is a function of the previous word (for a bi-gram language model) and the previous two words (for a tri-gram language model). Typically, the higher the order of the language model, the higher the recognition accuracy at the cost of slower recognition speeds. The language models may be trained on data that is manually and accurately transcribed by a human.

[0033] Speaker clustering logic 124 may identify all of the segments from the same speaker in a single document (i.e., a body of media that is contiguous in time (from beginning to end or from time A to time B)) and group them into speaker clusters. Speaker clustering logic 124 may then assign each of the speaker clusters a unique label. Speaker identification logic 125 may identify the speaker in each speaker cluster by name or gender.

[0034] Name spotting logic 126 may locate the names of people, places, and organizations in the transcription. Name spotting logic 126 may extract the names and store them in a database. Topic classification logic 127 may assign topics to the transcription. Each of the words in the transcription may contribute differently to each of the topics assigned to the transcription. Topic classification logic 127 may generate a rank-ordered list of all possible topics and corresponding scores for the transcription.

[0035] Story segmentation logic 128 may change the continuous stream of words in the transcription into document-like units with coherent sets of topic labels and other document features. This information may constitute metadata corresponding to the input audio data. Story segmentation logic 128 may output the metadata in the form of documents to memory system 130, where a document corresponds to a body of media that is contiguous in time (from beginning to end or from time A to time B).

[0036] In one implementation, logic 122-228 may be implemented in a manner similar to that described in John Makhoul et al., "Speech and Language Technologies for Audio Indexing and Retrieval," Proceedings of the IEEE, Vol. 88, No. 8, August 2000, pp. 1338-1353, which is incorporated herein by reference.

[0037] Memory system 130 may store documents from indexer 120. Memory system 130 may include one or more databases 131. Database 131 may include a conventional database, such as a relational database, that stores documents from indexer 120. Database 131 may also store documents received from clients 150 via server 140. Server 140 may include logic that interacts with memory system 130 to store documents in database 131, query or search database 131, and retrieve documents from database 131.

[0038] Server 140 may include a computer or another device that is capable of interacting with memory system 130 and clients 150 via network 160. Server 140 may receive queries from clients 150 and use the queries to retrieve relevant documents from memory system 130. Clients 150 may include personal computers, laptops, personal digital assistants, or other types of devices that are capable of interacting with server 140 to retrieve documents from memory system 130. Clients 150 may present information to users via a graphical user interface, such as a web browser window.

[0039] Typically, in the operation of system 100, audio streams are transcribed as rich transcriptions that include metadata that defines information, such as speaker identification and story segments, related to the audio streams. Indexer 120 generates the rich transcriptions. Clients 150, via server 140, may then search and browse the rich transcriptions.

[0040] Fig. 2 is a diagram illustrating an exemplary computing device 200 that may correspond to server 140 or clients 150. Computing device 200 may include bus 210, processor 220, main memory 230, read only memory (ROM) 240, storage device 250, input device 260, output device 270, and communication interface 280. Bus 210 permits communication among the components of computing device 200.

[0041] Processor 220 may include any type of conventional processor or microprocessor that interprets and executes instructions. Main memory 230 may include a random access memory (RAM) or another type of dynamic storage device that stores information and instructions for execution by processor 220. ROM 240 may include a conventional ROM device or another type of static storage device that stores static information and instructions for use by processor 220. Storage device 250 may include a magnetic and/or optical recording medium and its corresponding drive.

[0042] Input device 260 may include one or more conventional mechanisms that permit an operator to input information to computing device 200, such as a keyboard, a mouse, a pen, a number pad, a microphone and/or biometric mechanisms, etc. Output device 270 may include one or more conventional mechanisms that output information to the operator, including a display, a printer, speakers, etc. Communication interface 280 may include any transceiver-like mechanism that enables computing device 200 to communicate with other devices and/or systems. For example, communication interface 280 may include mechanisms for communicating with another device or system via a network, such as network 160.

EXEMPLARY PROCESSING

[0043] As previously mentioned, audio classification logic 221 classifies input audio streams into broad classes that can be used by logic 122-128 to improve speech recognition. Fig. 3 is a diagram illustrating functional components of audio classification logic 121. As shown, audio classification logic 121 includes signal analysis component 301 and decoder 302.

[0044] Signal analysis component 301 performs initial signal processing functions on the input audio stream. Signal analysis component 301 may convert the audio stream into a frequency domain signal and generate cepstral features for the audio. The conversion of an audio signal into the frequency domain and the calculation of cepstral features for an audio signal is well known in the art and will not be described in detail herein.

[0045] Decoder 302 receives the version of the audio stream processed by signal analysis component 301. Decoder 302 may generate an output signal that indicates the classification for a particular portion of the audio signal. The output signal may be a continuous signal or a signal that classifies the input audio signal in predetermined length segments (e.g., 10ms). As shown in Fig. 3, the classifications generated by decoder 302 may include a vowel class 310, a fricative class 311, a narrowband class 312, a wideband class 313, a coughing class 314, a gender class 315, and a silence class 316.

[0046] Vowel class 310 may include sounds uttered by a speaker that would traditionally be classified as one of a number of phonemes that define vowel and nasal sounds. Fricative class 311 may include sounds uttered by a speaker that that would be traditionally be classified as one of a number of phonemes that define obstruents and fricatives (consonants).

[0047] Narrowband class 312 and wideband class 313 relate to the bandwidth of the input audio signal. Narrowband signals may be signals received over a medium such as a telephone line or a radio broadcast. Wideband signals are higher quality signals, such as an audio signal received via a high-quality satellite broadcast. Speech recognition logic 123 may use different language models depending on whether the audio signal is a wideband or narrowband signal.

[0048] Coughing class 314 indicates a class for sounds that relate to non-speech events, such as a coughing sound. Other sounds that may be included in coughing class 314 includes laughter, breath and lip-smack sounds.

[0049] Gender class 315 indicates whether a sound is likely to have been produced by a male speaker or a female speaker. Silence class 316 indicates silence.

[0050] Decoder 302 may use a Hidden Markov Model (HMM) to model each of the seven phone classes. HMMs are probability models defined by a finite number of states. Transitions among the states are governed by a set of probabilities, called transition probabilities. HMMs are generally well known in the art.

[0051] Fig. 4 is a diagram illustrating decoder 302 in additional detail. Decoder 302 includes HMM based acoustic models for classifying the input audio streams. In particular, decoder 302 includes vowel model 401, fricative model 402, coughing model 403, silence model 404, narrowband model 405, wideband model 406, and gender model 407.

[0052] Vowel model 401, fricative model 402, and coughing model 403 may be phoneme based models. These models classify an audio signal based on rough phoneme classes. Vowel model 401 may, for example, determine if an input phoneme in the audio signal falls within what would conventionally be approximately 25 phonemes that relate to vowel sounds. Vowel model 401 does not need to determine which of the 25 conventional vowel phonemes corresponds to the input phoneme; only that the input phoneme falls within the vowel class of phonemes. Accordingly, vowel model 401 can be a less complicated and faster executing model than a model designed to determine the particular phoneme to which a sound corresponds.

[0053] Fricative model 402 and coughing model 403 may also be implemented as phoneme based models. In a manner similar to vowel model 401, these models determine whether sounds in the input audio signal correspond to any of the set of phonemes that correspond to these sounds.

Fricative model 402, for example, may include approximately 18 conventional phonemes that define consonant sounds.

[0054] Silence model 404, narrowband model 405, wideband model 406, and gender model 407 may be implemented based on non-phoneme properties of the audio signal. In particular, these models may be based on a conditional analysis of the signal from signal analysis component 301. Narrowband model 405 and wideband model 406 may examine the frequencies present in the signal. Signals with a bandwidth greater than 8 kHz may be classified as wideband signals while those having a bandwidth less than 8 kHz may be classified as narrowband signals. Additionally, signals with very little energy may be classified as silence by silence model 104. Gender model 107 may examine the frequency distribution of the signal to make a determination as to whether the speaker is a male or female.

[0055] As shown in Fig. 4, models 401-407 may be logically arranged such that vowel model 401, fricative model 402, coughing model 403, and silence model 404 initially process the input audio signal to classify the signal as belonging to one of the vowel, fricative, coughing, or silence class. Narrowband model 405 and wideband model 406 may then further classify the signal as being a narrowband or wideband signal. Gender model 407 may then further classify speech portions of the input signal as emanating from a male or a female speaker.

[0056] The output of decoder 302 may include a sequence of symbols that represent the classification results by models 401-407. Fig. 5 is a diagram illustrating an exemplary sequence of output symbols. In Fig. 5, "V" represents the vowel class, "F" represents the fricative class, "C" represents the coughing class, "S" represents the silence class, "W" represents a wideband signal, "N" represents a narrowband signal, "m" denotes a male gender classification, and "f" denotes a female gender classification. Symbol set 501 includes an exemplary series of five output symbols generated by the first stage of decoder 302 (i.e., models 401-404). Each of the symbols in set 501 may be associated with a start time within the input audio signal and a duration period. As shown, set 501 includes a vowel phoneme, a fricative phoneme, a vowel phoneme, a coughing sound, and silence. Symbol set 502 includes the symbols of set 501 after additionally being processed by models 405 and 406. The first symbol of symbol set 502, for example, is a vowel phoneme that is a narrowband signal. At the fourth symbol of symbol set 502, the signal switches to a wideband signal. Symbol set 503 includes the symbols of set 502 after additionally being processed by model 407. The first symbol of symbol set 503, for example, is a narrowband vowel phoneme that was spoken by a male.

[0057] Fig. 6 is a diagram illustrating training of a phoneme based model 601 consistent with an implementation of the invention. Model 601 may correspond to any one of models 401-403.

[0058] Model 601 may be trained using a recorded stream(s) of audio data that includes speech. The audio stream is passed through signal analysis component 301 and the resultant signal, labeled as input audio signal 610, is received by model 601. Additionally, model 601 receives a phone class sequence 612 (i.e., vowels, fricatives, or coughing) corresponding to the input signal 610. The phone class sequence 612 may be derived based on a meticulous manual transcription of the speech corresponding to input signal 610. Phoneme class dictionary 615 may be used to convert the transcribed words into the appropriate phoneme classes. Phoneme dictionaries that list the phonemes corresponding to a word are well known in the art. Phoneme class dictionary 615 may be implemented as a conventional phoneme dictionary that further converts the looked-up phonemes to their respective phoneme classes.

[0059] Model 601 may include HMMs. The HMMs are trained based on the input audio signal 610 and the phone class sequence 612 using conventional HMM training techniques.

[0060] Fig. 7 is a diagram illustrating training of a phoneme based model 701 consistent with another implementation of the invention. Model 701 may correspond to any one of models 401-403.

[0061] Model 701 may be trained using a recorded stream(s) of audio data that includes speech. Input audio signal 710 is similar to input audio signal 610. In particular, input audio signal 710 is an audio stream that contains speech data and is passed through signal analysis component 301. Additionally, as with model 601, model 701 receives a phone class sequence 712 generated by phone class dictionary 715 from transcribed speech. Phone class dictionary 715 is implemented identically to phone class dictionary 615.

[0062] In addition to the phoneme class sequence 712, model 701 receives information relating to the transcribed words, labeled as input 714. By receiving the transcribed words, as well as the phoneme class sequence 712, model 701 can determine which of the boundaries between the phone classes correspond to word boundaries. Stated differently, model 701 retains word boundary markings in the input phoneme classification, while model 601 simply receives a continuous stream of phoneme classifications. In some situations, this can lead to a more accurate model.

[0063] As with model 601, model 701 may include HMMs. The HMMs are trained based on the input audio signal 710, the phone class sequence 712, and the word boundaries derived from input words 714.

[0064] HMM based models 601 and 701 have been found to be trainable based on multiple languages.

CONCLUSION

[0065] The audio classification system, as described herein, classifies audio that includes speech data into a number of broad classes. Phoneme and condition models are used to perform the classifications. The classification information can be used to improve speech recognition.

[0066] The foregoing description of preferred embodiments of the invention provides illustration and description, but is not intended to be exhaustive or to limit the invention to the precise form disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention. For example, while seven broad classes were considered by decoder 302, in other implementations, more or fewer classes could be used.

[0067] Certain portions of the invention have been described as software that performs one or more functions. The software may more generally be implemented as any type of logic. This logic may include hardware, such as application specific integrated circuit or a field programmable gate array, software, or a combination of hardware and software.

[0068] No element, act, or instruction used in the description of the present application should be construed as critical or essential to the invention unless explicitly described as such. Also, as used herein, the article "a" is intended to include one or more items. Where only one item is intended, the term "one" or similar language is used.

[0069] The scope of the invention is defined by the claims and their equivalents.